DOCUMENT RESUME

ED 037 785 CG 005 190

AUTHOR Crawford, William R.
TITLE Assessing Performance When the Stakes are High.
INSTITUTION American Educational Research Association, Washington, D.C.; Illinois Univ., Urbana. Medical Center.
PUB DATE 2 Mar 70
NOTE 11p.; Paper presented at American Educational Research Association Convention, Minneapolis, Minnesota, March 2-6, 1970

EDRS PRICE EDRS Price MF-$0.25 HC-$0.65
DESCRIPTORS Achievement, *Achievement Rating, Achievement Tests, *Medical Students, *Performance Criteria, *Performance Tests, Simulation, Test Construction, *Testing Programs

ABSTRACT

This paper is concerned with measuring achievement levels of medical students. Precise tools are needed to assess the readiness of an individual to practice. The basic question then becomes, what can this candidate do, at a given time, under given circumstances. Given the definition of the circumstances, and the candidate's performance, the necessary evaluation of competence can be made. The University of Illinois College of Medicine has developed the "Minimum Passing Level" (MPL), which is a score point which reflects the minimum level of performance a student must achieve to progress at the normal rate or to be certified as eligible for the degree Doctor of Medicine. Both multiple choice examinations and simulated problems in patient management are used. A performance level must be established by examiners for each item seperately. The options are weighted on each item and given a weight from high negative to high positive. After the essential positive options and the not allowable negative options have been identified, the option scores are summed to determine the minimum required passing level. Disadvantages include: (1) subjective judgements, and (2) confusion regarding results and methods. (author/KJ)

ASSESSING PERFORMANCE WHEN THE STAKES ARE HIGH*

William R. Crawford
University of Illinois
at the Medical Center

As you may have inferred from the title of this paper and my institutional affiliation, my comments today will be directed primarily toward measuring achievement levels of medical students. However, I am convinced that the process I shall describe is applicable to other areas of education, in particular to professional education. Certainly, the stakes are especially high in medicine because of the immediate impact on the public and the potential for immeasurable good or irreparable harm.

Medical schools, like many other academic institutions, have for some time been trying to resolve issues concerning how levels of competence can be assessed best. Like other professional schools, they have traditionally placed great faith in those intangibles known as faculty insight and feelings and have used relative standards of performance to determine which students should be certified as competent. The time is long overdue for developing more precise tools for assessing the readiness of individuals to practice. In effect, there

---

*Prepared for an AERA symposium titled "Criterion-Related Measures: Bane or Boon?" Minneapolis, March 1970.

must be some method by which one can determine if it is safe to turn a potential practitioner loose on the public, whether he be a physician, nurse, or elementary school teacher. It is not enough to say that Dr. X is so much better than Dr. Y or that graduates of School A are so much better than those of School B when it is possible that in neither case do the individuals concerned have the basic minimal competencies necessary to practice their profession safely and efficiently. In comparing them with each other by norm-referenced methods, we risk losing sight of the far more important question of whether they can in fact do what it is they are supposed to be able to do. The basic question which should be considered in the assessment of professional competence then becomes: "What can this candidate do, at a given time, under certain circumstances?"

I do not intend to imply that all physicians must have exactly the same competencies, for to do so might be tantamount to asking for a cadre of automatons performing some functions which might be accomplished more easily and efficiently by a technician and a computer. Certainly, patient needs and requisite physician skills are different in various regions. There is abundant evidence to show that the needs in Berlin are not the same as those in the Belgian Congo and those in Vladivostok are not the same as those in Keokuk. Because our office

has been designated by the World Health Organization as one of two international centers (the other is in Moscow) we are acutely aware of the need for precise definition of physician skills in relation to patient needs in a given region. In fact, that need is precisely why the fundamental factor in our definition of criterion-related measure is: "What can this candidate do, at this point in time, under these circumstances?" Given the definition of the circumstances and the candidate's performance the necessary evaluations of competence can be made.

For several years the Interdepartmental Appraisal Committee of the University of Illinois College of Medicine has addressed itself to the question of developing meaningful definitions of appropriate performance levels. Over the past eight years the College has developed an intricate system of comprehensive examinations and teaching-learning examinations (Crawford, 1966; McGuire, 1967). The component of the examination program which is of primary interest today is the technique and concept which we call the "Minimum Passing Level."

Basically, the "Minimum Passing Level," which I shall refer to as MPL, is a score point which reflects the minimum level of performance a student must achieve to progress at the normal rate or to be certified as eligible for the degree Doctor of Medicine.

As conceived by the Committee it can be applied to many aspects of the curriculum including knowledge acquisition, problem-solving skills, and clinical skills in an office or hospital setting. At the present time, it is applied most strictly in the assessment of cognitive achievement. These aspects of expected student behavior are measured with two types of paper-and-pencil tests; multiple choice examinations and simulated Problems in Patient Management. The limitation of today's discussion to the cognitive aspects of behavior does not imply that the faculty believes it is the sole factor which is of importance. They are acutely aware of the importance of attitudes, values, and skills, but time does not permit discussion of those characteristics.

Let us look first at the technique used to establish the MPL in the multiple choice portion of the comprehensive examinations. The method currently in use is similar to the one described by Nedelsky (1954). It requires that the examiner scrutinize each alternative to every item and determine which alternatives a barely satisfactory student should be able to eliminate from consideration, or considering it the other way around, between which alternatives might the barely passing student experience difficulty in discriminating. An alternative procedure favored by some examiners produces similar results. In it <u>each item</u> is assigned a value based on an estimate

of how well a student should be expected to perform on a large number of similar items. Through this process the examiner arrives at a most likely performance estimate for each item in the examination. These expected item performance levels are then combined to obtain the most likely total score for a hypothetical barely satisfactory student. It is critical to note that a performance level must be established by examiners for each item separately without anticipating the total which will result. This procedure, and an additional safeguard to be mentioned later, helps minimize the possibility that the examiner might establish an arbitrary overall criterion of, say, 75% = passing, and work backward to establish values for the items.

The MPL's for Problems in Patient Management are established in a slightly different way. Time does not permit me to describe the unique characteristics of Patient Management Problems, but they have been described elsewhere. (Charvat, McGuire & Parsons, 1968; McGuire & Babbott, 1967; Williamson, 1965).

Briefly, the problems are designed to simulate clinical physician-patient encounters and are based on a principle of branched design making sequences so that every student need not follow the same path through a problem. Because of the branching, alternative decision feature it was necessary to develop a new procedure for defining

the acceptable level of criterion-related performance. The procedure is based on the concept of what a physician must do and what he must avoid for a given patient, at a given point in time, under carefully defined circumstances. To determine these factors it was necessary to define two basic types of errors which we call "errors of omission" and "errors of commission." Each option in a problem is carefully examined to determine if it is an essential step which must be taken to assure good management in a specific situation. If an option so identified is not chosen by a student he has committed an error of omission. Conversely, if he chooses an option identified as being detrimental to the welfare of the patient the student has made an error of commission. The options are weighted on a scale which ranges from high negative through zero to high positive depending on the relative detriment or benefit the option represents. After the essential positive options and the not-allowable negative options have been identified the option-score-values are summed to determine the maximum permissible negative score and minimum required positive score. These values are then summed algebraically to determine the net score which a minimally competent student must achieve. This minimum net score is the Minimum Passing Level.

The initial assignment of MPL values is not the final step in the determination of acceptable performance levels. Complete item

and test data are obtained by computer (Lewy & Crawford, 1966) and each item is reviewed in light of the item data. It is of substantial significance that a this point the examiners have no knowledge of the distribution of scores for either the content areas or the total score. They review the items solely on the basis of item data and, if necessary, revise the MPL or delete the item from scoring. This provides the additional safeguard against working from test to item data that I mentioned earlier. On the basis of this review a new total MPL is calculated if necessary and the examination is rescored omitting items in which the examiners may have found flaws which were not detected in the initial review. I should emphasize that the changes instituted at this point are minimal indeed and in most cases do not result in a significant change in the total MPL or the MPL's for specific content areas.

Clearly, the procedure which I have described involves some subjective judgment in determining the multiple-choice alternatives and Patient Management Problem options which are essential and those which should or must be avoided. As you may guess, this is the criticism of the technique which is most often heard. However, I submit that this procedure is substantially less subjective than a post-hoc arbitrary assignment of cut-off points based on centile rank, standard score point, or percent of correct responses, particularly when these techniques are applied to highly select or homogeneous groups. Further,

our experience has been that faculty, working independently, exhibit a high degree of rater agreement in reaching these decisions and that even when discrepancies in ratings do occur they can generally be resolved with little difficulty. One of the beneficial side-effects of this procedure is that it forces the faculty to examine their items much more carefully than other systems of which we are aware. Further, this technique reveals genuine changes over time in students, curriculum, and teaching that are obscured by norm-referenced procedures.

It is by no means necessary to base the item and option ratings solely on subjective evaluations established by the faculty. In cooperation with the American Board of Orthopaedic Surgery Professor McGuire, Mr. Harold Levine, and Dr. Carl Olson have constructed and studied criterion-related certification examinations based on extensive task analysis and critical incident findings (Levine, McGuire, Miller & Larson, 1968). Similarly, through a grant from the Illinois Regional Medical Program, Mrs. Merry Omori and I are beginning to study patient needs in relation to the availability of health care in the State of Illinois. We anticipate that a series of criterion-related measures will result from the investigations which will be of particular value to physician self-assessment efforts and to continuing education programs.

Another criticism sometimes leveled at proponents of criterion-related measurement techniques is that we advocate mastery tests,

i.e., tests on which students should achieve with about 95-100% accuracy. Such is simply not the case. There is nothing in the definition of criterion-related measures which prohibits the use of items at varying levels of difficulty. In fact, the procedures I have outlined today depend to some extent on the existence of variance.

Basically, the position we take is that the examination should be constructed to measure the critical components of behavior and the item writer must not be ham-strung by a statistical model. It is possible, and indeed has happened, that on some of our examinations the MPL will be as low as 35% and on others as high as 80%, and the failure rate in principle can vary from 0-100% and <u>has</u> varied from 0-90%. Remember, these are minimum levels, not optimum levels. It is possible, of course, to define the minimum score for superior achievement in the same way it is defined for barely satisfactory performance. I did not discuss this aspect because our major concern to date has been in developing a procedure for identifying students who have failed to achieve the required levels of professional proficiency. Our students do receive rewards for outstanding performance, but this procedure is a direct responsibility of the Office of the Dean.

As Professor Hills has mentioned, numerous problems are encountered when one attempts to interpret classical item and score data which

have been derived from criterion-related measures. For those of us schooled in the classical approaches to difficulty, discrimination, reliability, and validity, the results from criterion-related measures can at first be a little confusing. Nevertheless, with experience one begins to establish new criteria for evaluating these data which probably are at least as objective and accurate as the classical "rules of thumb." If you take the no-variance approach to criterion-related measurement, the problems become more acute because without variance our classical statistical models become meaningless. I believe the no-variance approach is neither necessary nor desirable in most instances.

The procedures and topics I have discussed here must be considered in the context of the type of institution in which they are applied. It may be true that these procedures are easier to apply and are of more critical importance in medical education than in some other, less precisely defined areas, but so far I am not convinced of that. I would like to add that I have used similar procedures in my own courses in educational psychology with success. However, the success was achieved only after substantial initial student resistance was overcome. I believe it was something of an accomplishment when you consider that this resistance was encountered on the campus of San Francisco State College and overcome without the protection of a flak helmet.

References

Charvat, T., McGuire, C., & Parsons, V. A Review of the Nature and Uses of Examinations in Medical Education. Geneva: World Health Organization 1968.

Crawford, W. R. A validation of the structure and generality of "A Taxonomy of Intellectual Processes." (Doctoral dissertation, The Florida State University, Tallahassee) Ann Arbor, Michigan: University Microfilms 1966, No. 66-9063, pp. 14-35.

Levine, H. G., McGuire, C., Miller, G. E., & Larson, C. The Orthopaedic Training Final Report. (Research Grant No. PM 00014, Bureau of Health Manpower, Public Health Service, Department of Health, Education and Welfare). Chicago, University of Illinois at the Medical Center 1968.

Lewy, A. & Crawford, W. R. Scoring test battery: A program for the IBM 7094. Educational and Psychological Measurement, 1966, 26, 185-188.

McGuire, C. An Evaluation Model for Professional Education, Invitational Conference on Testing Problems. Princeton, New Jersey, Educational Testing Service, 1967, pp. 37-52.

McGuire, C. & Babbott, D. Simulation technique in the measurement of problem-solving skills. Journal of Educational Measurement, 1967, 4, 1-10.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Williamson, J. W. Assessing clinical judgment. Journal of Medical Education, 1965, 40, 180-187.